REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
UNCLASSIFIED	

2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	Approved for public release; distribution is unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
NPRDC TN 88-37	

6a. NAME OF PERFORMING ORGANIZATION	6b OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Navy Personnel Research and Development Center	Code 62	

6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
San Diego, CA 92152-6800	

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
Chief of Naval Research	ONT	Washington, DC 20350

8c. ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO
Washington, DC 20350	62233N	RM33M20	4a	

11 TITLE (Include Security Classification)
Reliability and Construct Validity of Reaction Time, Inspection Time, and Machine-paced Tests of Cognitive Speed

12. PERSONAL AUTHOR(S)
G. E. Larson, C. R. Merritt, & K. E. Lattin

13a. TYPE OF REPORT	13b TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Note	FROM FY87 TO FY87	1988	16

16 SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Cognitive speed, reaction time, inspection time, information processing, intelligence
05	09		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In order to evaluate the psychometric characteristics of cognitive speed tests, a battery of reaction time, inspection time, and machine-paced tests was administered to 267 male Navy recruits. Two hundred and twenty of these subjects returned approximately one month later to be retested. Our final evaluation of each test was based on whether or not it met two or more of the following standards: Test-retest reliability of .70, split-half reliability of .90, and construct validity of .30. The results indicate that the majority of the tests meet these standards.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Gerald E. Larson	(619) 553-7656	Code 62

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

Navy Personnel Research and Development Center

San Diego, CA 92152-6800   TN 88-37   April 1988

Reliability and Construct Validity of Reaction Time, Inspection Time, and Machine-paced Tests of Cognitive Speed
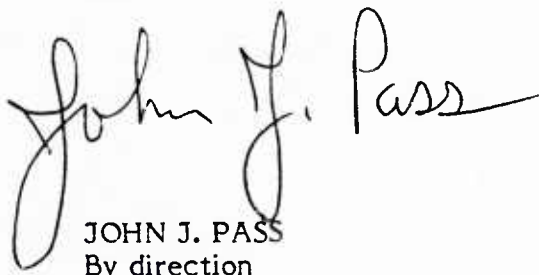
3900
Ser 62/361
26 APR 1988

From: Commanding Officer, Navy Personnel Research and Development Center

Subj: **RELIABILITY AND CONSTRUCT VALIDITY OF REACTION TIME, INSPECTION TIME, AND MACHINE-PACED TESTS OF COGNITIVE SPEED**

Encl: (1) NPRDC TN 88-37

1. The present research is an outgrowth of the Cognitive Speed project at the Navy Personnel Research and Development Center (NAVPERSRANDCEN). The project's purpose was to determine whether tests of mental speed could be used to supplement information provided by the current Armed Services Vocational Aptitude Battery. Cognitive speed measures may provide a broader ability profile on which personnel selection and classification could be based. The work is being continued under the Personnel Performance Prediction project (Work Unit Number 88WX4B529) in FY 1988.

2. Earlier research at NAVPERSRANDCEN indicated that speed of information processing is related to mental aptitude (NPRDC TR 86-23) and performance in a technical training program (NPRDC TR 85-3). Prior work on civilian college students also suggests that cognitive speed can be reliably measured (NPRDC TN 88-10). The present report describes reliability for a military sample.

JOHN J. PASS
By direction

Reliability and Construct Validity of Reaction Time, Inspection Time, and Machine-paced Tests of Cognitive Speed

Gerald E. Larson
Charles R. Merritt
Kathryn E. Lattin

Approved by
John J. Pass, Ph.D.
Director, Personnel Systems Department

SUMMARY

Problem

Cognitive speed, or the quickness with which individuals can interpret and/or respond to information, is of interest as a possible supplement to skills measured by the Armed Services Vocational Aptitude Battery (ASVAB). Before cognitive speed tests can be used for selection and classification purposes, however, adequate test-retest reliabilities must be demonstrated. The issue is critical since reliability reflects the extent to which performance on a test is biased by irrelevant factors, including chance.

Purpose

The purpose of the present investigation was to determine the reliability (in particular, test-retest) coefficients for a battery of cognitive speed tests. In addition, practice effects and construct validity were explored.

Approach

A battery comprised of three reaction time (RT) tests, two tests involving machine-paced (MP) item frames, and an inspection time (IT) test was administered to 267 male Navy recruits. Two hundred and twenty subjects returned 4 weeks later for retesting. In addition, a nonverbal test of intelligence (Raven's Advanced Progressive Matrices) was administered, and scores on the Armed Forces Qualification Test (AFQT) were gathered from the subject's personnel records. The Raven and AFQT were averaged to form a general intelligence score.

Results and Discussion

Reliabilities

Using .70 as the standard of adequate reliability for tests undergoing continued development, only Simple RT and one of the scores from the Arrows RT test had adequate test-retest coefficients. There is a strong possibility, however, that some of the other tests are reliable enough to contribute to personnel measurement as part of a test composite. Composites based on RT and MP paradigms, respectively, have test-retest reliabilities greater than .70. Composites in which paper and pencil test scores are combined with the computerized tests were not developed for the current study, but are a promising area for future research.

Using .90 as a standard, Simple RT, Arrows RT, and the two MP tests (Numbers and Counters) have split-half reliabilities high enough to justify further research and development.

Practice Effects

The subjects in the study improved with practice on all tasks except those that were largely perceptual (i.e., IT, Simple RT, and the Physical Identity (PI) test). Stated differently, tasks that require mental operations upon test items are also those that benefit from practice. This improvement could presumably result either from the acquisition of strategies or the general automaticity of cognitive operations following practice. However, the present study was not designed to clarify the manner in which repeated exposure to the items used in the study led to improvement.

Construct Validity

Of the computerized tests, only the two MP tests (Counter and Numbers) had correlations with general intelligence higher than .30, and thus adequate construct validity. The distinguishing feature of the MP tests is that they involve more sustained mental effort, and, particularly for Counters, greater complexity/momentary workload than the other experimental tests in the study. This finding is supported by other research indicating that task complexity is largely responsible for the magnitude of correlations between single tests and a general factor of intelligence (Marshalek, Lohman, & Snow, 1983). An important goal for future research is to determine exactly how task complexity/workload affects the accuracy with which a test differentiates between individuals of high and low ability.

Recommendations

The present study evaluated a series of cognitive speed tests on several dimensions. We now consider all these dimensions together in judging the potential of the tests. In our final evaluation, we recommend further research and development on tests that equalled or exceeded at least two of the three previously cited standards (e.g., test-retest reliability of .70, split-half reliability of .90, and construct validity of .30). The recommended tests are Mental Counters, Numbers, Arrows RT, and Simple RT. The following tests do not appear to warrant further development: IT and the two letter matching RT tasks (Name Identity (NI) and PI).

CONTENTS

LIST OF TABLES

INTRODUCTION

The present report is the second to document the reliabilities of a set of cognitive speed tests developed by the Navy Personnel Research and Development Center (NAVPERSRANDCEN). Cognitive (or mental) speed, defined here as the quickness with which individuals can interpret and/or respond to information, is of interest as a possible supplement to skills measured by the Armed Services Vocational Aptitude Battery (ASVAB). There are three main paradigms for measuring speed: (1) Reaction time (RT), where a subject must press a response key as quickly as possible following presentation of a test item. (2) Inspection time (IT), where subjects must try and correctly perceive, in accurate detail, simple displays that are presented for only a fraction of a second. (3) Machine-paced (MP), where sequential video frames are rapidly shown to subjects who must process information at the same rate as the display sequence.

In an earlier report, Saccuzzo and Larson (1987) evaluated various tests of the RT, IT, and MP paradigms and noted, for each paradigm, the tests that appeared most reliable and/or practical. In the present research, those core tests are reevaluated in order to verify the reliabilities reported by Saccuzzo and Larson (1987). Secondary goals of the present research were to examine practice effects for the various paradigms and to determine the construct validity of cognitive speed tests.

METHODS

Subjects

Subjects were male Navy recruits (N = 267; mean age 19.8 years) selected at random from groups undergoing in-processing at the Recruit Training Command, San Diego. Subjects were tested twice on a battery of cognitive speed measures with approximately one month separation between sessions. Due to scheduling conflicts, however, only 220 of the subjects were able to return for the retest. There is no reason to suspect that the "drop-outs" were nonrandom with respect to the cognitive ability variables in the study, given the nature of their schedule conflicts (dental appointments, swimming lessons, etc.).

Cognitive Speed Tests

Each subject was administered a battery of computerized tests, presented on IBM PC/XT microcomputers with color monitors and standard keyboards. No special add-ons were used other than color labeling of response keys and anti-glare filters for the monitors. Order of test presentation was completely randomized for each subject according to a prearranged sequence.

1. Reaction Time Paradigms

Three RT tests were administered: (a) Simple RT, (b) Arrows, and (c) a letter matching task.

a. Simple RT

A .25 inch open square in the center of the cathode ray tube (CRT) screen was used as a stimulus. Subjects were instructed to respond by pressing the space bar as quickly as possible after the square became illuminated. At the beginning of each trial, the subject rested the forefinger of his dominant hand on the space bar at the bottom of

the keyboard. After periods of from 1 to 6 seconds, the square was illuminated. RT was the number of milliseconds between stimulus onset and the instant the subject pressed the space bar. There were 80 trials. If a RT greater than 2 seconds was recorded, the trial was discarded and a new one presented to maintain a total of 80. A count was kept of discarded trials. Median RT was used in subsequent analyses, along with the standard deviation of the "good" trials.

b. Arrows

In the Arrows test (Larson, 1986), subjects were instructed to fixate on two small circles (the lowercase letter "o") presented 0.5 inches apart in the center of the CRT screen. For each trial, one of the circles was replaced by an arrow, and, depending on the arrow's direction and position, the subject responded by pressing either a right or left key on the microcomputer keyboard. If the arrow pointed down, its position indicated the appropriate response. For example, if a down-pointing arrow replaced the right circle, the right key was pressed. If an arrow pointing right or left was presented, then its direction became the relevant cue while position became a distractor. For example, if an arrow appearing on either side pointed right, the right key was pressed. The position and direction of the arrow were varied randomly. The test involved 82 trials; 41 with downward arrows and 41 with right-left arrows. RTs greater than 2 seconds were discarded and new items presented to maintain a constant number of trials per subject. A count was kept of discarded trials. Median response latencies for downward (ARROW-DOWN) and right-left (ARROWSIDE) arrows were included in subsequent analyses, along with an overall standard deviation.

c. Letter Matching

The letter matching tasks were based on the work of Posner and Mitchell (1967). There were two subtests—Physical Identity Test (PI) and Name Identity Test (NI). In the PI test, subjects were required to make judgments based on the physical appearance of two letters. For example, the letters "a" and "a" look the same, whereas, the letters "A" and "a" or "g" and "d" look different. Response times for same and different judgments were recorded for each trial. In the NI test, subjects must respond based on the names of two letters. For example, the letters "a" and "A" have the same name, while "a" and "c" do not.

On both tests, subjects were instructed to fixate on a period (".") located in the center of the screen. Following a random wait of 1.5 to 2.5 seconds, the period was replaced by two letters and the latency and accuracy of the subject's response were recorded. RTs greater than 2 seconds were discarded and new items presented to maintain a constant number of trials per subject. A count was kept of discarded trials. Each test consisted of 34 trials. The PI test was always presented first.

2. Inspection Time (IT)

In this task, subjects were briefly shown two horizontal lines of unequal length, presented in the center of the CRT screen. For each trial, the task was to determine which line in the pair is longer. Immediately following stimulus termination, a backward visual noise mask was presented. The mask is known to limit the duration of the sensory signal delivered to the central nervous system (Felsten & Wasserman, 1980). The subject's task was to make a forced-choice discrimination by pressing one of two keys on the microcomputer keyboard. Final score was the total number of correct responses. Following each response, subjects were given computer-generated performance feedback.

Test display is terminated by a backward mask in the form of a spatially overlapping line that obscures the test item and limits viewing time. Stimulus duration is the chief source of item difficulty. Five stimulus durations were used: 16.7, 33.4, 66.8, 100.2, and 150.3 milliseconds (ms). There were 15 trials per duration, presented in a prearranged random sequence, for a total of 75 trials. The lengths of the test lines were 17.5 mm and 14.3 mm. The distance between the lines was randomly determined for each trial, but ranged from 6 to 20 mm.

3. Machine-Paced (MP) Paradigms

MP tests present the subject with a rapid series of video frames, with each frame containing information critical to ongoing cognitive operations. To be successful, subjects must be able to process information at the rate the frames are presented. Two MP tests were administered: (a) Mental Counters and (b) Numbers.

a. Mental Counters

In the Mental Counters test (Larson, 1986), subjects must keep track of the values of three independent "counters," which change rapidly and in random order. (The difficulty of the task comes from having to simultaneously hold, revise, and store three counter values under severe time pressure. Slow execution of counter adjustments leads to a general breakdown on the task.) The counters are represented as lines on the video monitor (three side-by-side 1.0 inch horizontal dashes in the center of the screen). The initial counter values are zero. When a small target (a .25 inch box) appears above a dash, the corresponding counter must be adjusted by adding "1." When the target appears below one of the three dashes, the corresponding counter must be adjusted by subtracting "1," (see Figure 1). The test items vary in the number of adjustments and the rate of presentation. There were two levels of counter adjustments (five and seven) and two levels of rate of presentation (fast and slow). The actual test involved a total of 40 trials. On 20 trials, 5 adjustments were required. Seven adjustments were required on the remaining 20 trials. On 20 trials, adjustments were required at the rate of one every .75 seconds. On the remaining 20 trials, adjustments were required at the rate of one every 1.33 seconds. Number of targets and rate of presentation were completely counter-balanced. Total correct was used as the summary score.

b. Numbers

The Numbers test is a modification of a pre-probed digit encoding task described by Cohen and Sandberg (1980). In our version of the test, subjects were given a target digit to remember. Subjects were then asked to observe a rapidly presented sequence of 30 digits (shown one at a time), which included a single instance of the target in the middle third of the sequence (e.g., from serial position 9 to position 19 in the sequence). Following the presentation, the subject's task was to report the number shown before the target digit, the target, and the number after the target. The response was scored as "right" if all three numbers were reported in the correct order, and "wrong" if otherwise. The test was divided into two blocks of 20 trials each. In the first block, digits were presented at a rate of one every .43 seconds. In the second block, digits were presented at a rate of one every .26 seconds. Total correct across blocks was used as the summary score.

3

STEP	WHAT THE SUBJECT SEES	COUNTER ADJUSTMENT	COUNTER VALUES
0	___ ___ ___	None	0 0 0
1	☐ ___ ___	+1 x x	1 0 0
2	___ ☐ ___	x +1 x	1 1 0
3	___ ☐ ___	x -1 x	1 0 0
4	☐ ___ ___	+1 x x	2 0 0
5	___ ___ ☐	x x -1	2 0 -1

Please select your answer:

    1.   2   0   0

    2.   2   0 -1   (Correct answer is #2).

    3.   1   0 -1

    4.   2   1 -1

Figure 1. Sample item from Mental Counters test.

General Intelligence Tests

Armed Forces Qualifying Test (AFQT) scores were gathered from the recruits' personnel records. The AFQT, which is a composite of verbal and quantitative subtests, is used by the Armed Forces as a measure of general intellectual aptitude/trainability. In addition, the Raven Progressive Matrices (RPM) Test, Advanced (Raven, 1962) was administered with a 40 minute time limit. The Raven is a nonverbal test designed as a measure of general intelligence.

Our preliminary analyses indicated that the AFQT and the RPM were significantly correlated in the present sample ($r = .52$, $p < .01$). Based on the nature of the tasks, this finding appears to support the argument that verbal or knowledge-based tests and nonverbal measures of reasoning assess two correlated aspects of general intelligence (e.g., Cattell, 1971). Thus, we combined the verbal/quantitative AFQT and the nonverbal RPM into a general IQ score, which will be used to represent the construct of general intelligence in the analyses that follow.

The main questions to be addressed in the study are whether:

1. The test-retest and split-half reliabilities of the cognitive speed tests are high enough to justify further research and development.

2. There are practice effects for cognitive speed tests.

3. Construct validity (e.g., correlations with general intelligence) varies as function of cognitive speed paradigm (RT, IT, and MP).

RESULTS

Reliabilities

Test-retest

As Nunnally (1967) indicates, what a satisfactory level of reliability is depends on how a test is used. He suggests that in the early stages of research on predictor tests, reliabilities of .70 or higher are adequate. Although we adopt .70 as a goal for tests undergoing continued development, we also qualify this by noting two reasons why a rigid standard is inappropriate. First, since some of the testing technologies we used are highly experimental, much test refinement can still be anticipated. Second, test composites are typically used for personnel decisions within the Armed Forces. The reliability/validity of a composite score can be good even when the reliabilities of its components are modest. As Guilford and Fruchter (1973) note, "Tests with reliability coefficients as low as .35 have been found useful when utilized in batteries with other tests" (p. 91). There is a risk, then, of discarding a test with modest reliability that would have made a valuable contribution as part of a composite score.

Table 1 shows means and standard deviations for the first and second sessions on the computerized tests, along with test-retest reliabilities. The only scores to meet the .70 reliability standard are Simple RT and one of the medians from the Arrows RT test. Most of the other reliabilities are in the mid .60s.

Table 1

**Means, Standard Deviations, and Reliability Coefficients
for Computerized Tests**

Test	M1	M2	SD1	SD2	rxx	Saccuzzo rxx
IT	53.42	53.76	6.81	7.21	.64*	.66
Reaction Time						
Simple RT	312.54	335.98	45.65	59.59	.70*	.62
ARROWDOWN	534.39	526.48	91.15	96.63	.65*	.73
ARROWSIDE	593.87	569.48	98.30	109.45	.71*	.81
PI	621.66	623.47	90.96	113.61	.61*	--
NI	777.20	759.40	114.48	132.14	.61*	--
Machine-Paced						
Counters	27.24	28.72	7.62	7.83	.64*	.59
Numbers	21.65	23.31	7.20	7.74	.66*	--

*$p < .01$.

Where possible, test-retest reliabilities reported by Saccuzzo and Larson (1987) are also included in Table 1 for the sake of comparison. Several differences between the Saccuzzo and Larson study and the one reported here should be noted, however. First, his subjects were retested within one week. Test sessions for the present subjects were approximately one month apart. Second, Saccuzzo and Larson used an inspection time test with 50 trials, while the present version involved 75 trials. Finally, Saccuzzo and Larson's Simple RT task involved 21 trials, while the present version included 80 trials.

Normally, studies with longer time intervals between test sessions report lower reliability coefficients. All things being equal, our longer retest intervals should thus have produced reliabilities lower than those reported by Saccuzzo and Larson. That was indeed the case for the Arrows RT test. The higher reliabilities for IT and Simple RT in the present study are probably the result of using lengthened versions of these tests. The reason for the higher reliability of Counters, relative to the Saccuzzo and Larson study, is unknown.

Split-half

Split-half reliabilities for the computerized tests (first session data) are presented in Table 2. The analyses are based on the full sample. Using .90 as a standard, Simple RT, Arrows RT, and the two MP tests (Numbers and Counters) have split-half reliabilities high enough to justify further developmental work on these tests.

Table 2

Split-half Reliability Coefficients

Test	rxx
IT	.75
Reaction Time	
Simple RT	.94
Arrows	.92
PI	.87
NI	.85
Machine-Paced	
Counters	.93
Numbers	.91

N = 271.

Reliability of Test Composites

As noted above, test composites, which are commonly used by the Armed Forces, are often more reliable than the tests they incorporate. To determine whether reliable (and interpretable) composites could be built from the measures in our study, we factor-analyzed the scores from the computerized tests. The results are shown in Table 3. Three factors emerged in the varimax rotated matrix, interpretable as RT, MP, and IT paradigms, respectively. Since IT was represented by a single task, no composite could be formed. The five RT and two MP paradigms, however, were grouped into RT and MP composites, respectively. The test-retest reliability of the RT composite was .80. The MP composite had a reliability of .71. Overall, the test-retest data indicate that the constructs of RT and MP can be reliably measured, but that further developmental work on individual tests may be needed.

Practice Effects

Table 4 provides a summary of tests of significance for the difference between the means on first and second testings (first session minus second session). A sign difference for practice effects should be noted: Because RT scores are latencies, a positive difference score indicates improvement (e.g., faster responding). Since IT and MP scores are accuracy-based, a negative difference score indicates improvement (e.g., more correct on the second session). The subjects improved significantly on all tests except IT, PI, and Simple RT. An unexpected finding was that subjects were significantly worse (e.g., slower) on the second administration of Simple RT.

Overall, the results appear to indicate that the "non-cognitive" tasks were the ones that did not show improvement. IT and Simple RT are both primarily measures of perceptual encoding, in that neither test requires cognitive operations and/or mental transformations of test items. The PI test does require that a choice between stimuli be made, but the response is based only on the visual similarity of items. In the NI test, on the other hand, names retrieved from memory are the basis for the response choice. Data

for the NI test indicate significant improvement with practice. In conclusion, the only tasks that benefited from practice were those that involved some nontrivial degree of mental work. There was no clear relationship, however, between the amount of mental work required by a test and the degree of improvement on that test.

Table 3

Varimax Rotated Factor Matrix for Computerized Tests

Test	Factor 1	Factor 2	Factor 3
IT	-0.08	0.14	0.91
Reaction Time			
Simple RT	0.50	-0.17	-0.02
ARROWSIDE	0.89	-0.02	-0.17
ARROWDOWN	0.90	0.01	-0.22
NI	0.64	-0.43	0.23
PI	0.74	-0.37	0.20
Machine-Paced			
Counters	-0.11	0.75	0.21
Numbers	-0.15	0.82	-0.00

Table 4

Summary of t-tests for First and Second Testings

Test	Difference Between Means	SD	T-Value	Df	2-Tail Prob.
IT	-.34	5.92	-.85	219	NS
Reaction Time					
Simple RT	-23.43	42.83	-7.50	187	$p < .01$
ARROWSIDE	11.91	78.86	2.08	188	$p < .05$
ARROWDOWN	24.39	79.37	4.22	188	$p < .01$
PI	-1.82	92.76	-.28	208	NS
NI	17.79	110.41	2.32	207	$p < .05$
Machine-Paced					
Counters	-1.48	6.53	-3.29	210	$p < .01$
Numbers	-1.66	6.22	-3.81	203	$p < .01$

Construct Validity

There has been widespread speculation that speed of information processing is a fundamental basis of individual differences in intelligence (e.g., Jensen, 1982; Brand & Deary, 1982). Thus, correlations with standard tests of intelligence can be considered a form of construct validity for cognitive speed tests. As noted in the method section, we combined the Raven and AFQT into a composite paper and pencil measure of general intelligence. Correlations between this intelligence score, and scores from the computerized tests, are shown in Table 5. We use .30 as a standard for adequate construct validity, since .30 has been cited as a typical upper bound in speed/intelligence correlations (e.g., Hunt, 1980). The two MP computerized paradigms, Mental Counters and the Numbers test, had good correlations with general intelligence (.55 and .36, respectively, $p < .01$). More standard RT and IT paradigms, by contrast, had low to modest correlations with the composite intelligence score, and thus poor construct validity.

Table 5

**Correlations of Computerized Tests
with General Intelligence**

IT	.23**
Reaction Time	
Simple RT	-.16**
ARROWSIDE	-.13*
ARROWDOWN	-.14*
PI	-.22**
NI	-.24**
Machine-Paced	
Counters	.55**
Numbers	.36**

*$p < .05$.
**$p < .01$.

The distinguishing feature of the MP tasks is that they involve more sustained mental effort, and, particularly for Counters, a greater momentary mental workload than the other experimental tests in the study. An understanding of how test workload/complexity determines the usefulness of a test as a measure of intelligence is undoubtedly a matter of great theoretical importance.

DISCUSSION AND CONCLUSIONS

Reliabilities

Using .70 as the standard for adequate test-retest reliability for tests undergoing continued development, only Simple RT and one of the scores from the Arrows RT test had adequate test-retest coefficients. There is a strong possibility, however, that some of

the other tests with good construct validity are reliable enough to contribute to personnel measurement as part of a test composite. Composites based on RT and MP paradigms, respectively, have test-retest reliabilities greater than .70. Composite in which paper and pencil test scores are combined with the computerized tests were not developed for the current study, but are a promising area for future research.

Using .90 as a standard, Simple RT, Arrows RT, and the two MP tests (Numbers and Counters) have split-half reliabilities high enough to justify further research and development.

Practice Effects

The subjects in the study improved with practice on all tasks except those that were largely perceptual (i.e., IT, Simple RT, and the PI test). Stated differently, tasks that require mental operations upon test items are also those that benefit from practice. Improvement on the more complex tasks could presumably result from the acquisition of strategies or the general automaticity of cognitive operations following practice. The present study was not designed, however, to clarify the manner in which repeated exposure to the items used in the study led to improvement.

Construct Validity

Of the computerized tests, only the MP tests (Counters and Numbers) had correlations with general intelligence greater than .30, and thus adequate construct validity. The distinguishing feature of the MP tests is that they involve more sustained mental effort, and, particularly for Counters, greater complexity/momentary workload than the other experimental tests in the study. This finding is supported by other research indicating that task complexity is largely responsible for the magnitude of correlations between single tests and a general factor of intelligence (Marshalek, Lohman, & Snow, 1983). An important goal for future research is to determine exactly how task complexity/workload affects the accuracy with which a test differentiates between individuals of high and low ability.

RECOMMENDATIONS

The present study evaluated a series of cognitive speed tests on several dimensions. We now consider all these dimensions together in judging the potential of the tests. In our final evaluation, we recommend further research and development on tests that equalled or exceeded at least two of the three previously cited standards (e.g., test-retest reliability of .70, split-half reliability of .90, and construct validity of .30). The recommended tests are Mental Counters, Numbers, Arrows RT, and Simple RT. The following tests do not appear to warrant further development: IT and the two letter matching RT tasks (NI and PI).

REFERENCES

Brand, C. R., & Deary, I. J. (1982). Intelligence and "inspection time." In H. J. Eysenck, Ed.), A model for intelligence (pp. 93-132). New York: Springer-Verlag.

Cattell, R. B. (1971). Abilities: Their structure, growth, and action. New York: Houghton Mifflin.

Cohen, R. L., & Sandberg, T. (1980). Intelligence and short-term memory: A clandestine relationship. Intelligence, 4, 319-331.

Felsten, G., & Wasserman, G. S. (1980). Visual masking: Mechanisms and theories. Psychological Bulletin, 88, 329-353.

Guilford, J. P., & Fruchter, B. (1973). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hunt, E. (1980). Intelligence as an information processing concept. British Journal of Psychology, 71, 449-474.

Jensen, A. R. (1982). Reaction time and psychometric "g." In H. J. Eysenck, (Ed.), A model for intelligence (pp. 93-132). New York: Springer-Verlag.

Larson, G. E. (1986). The Arrows Test. Submission to the Joint-Service Future Testing Committee.

Larson, G. E. (1986). The Mental Counters Test. Submission to the Joint-Service Future Testing Committee.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. Intelligence, 7, 107-127.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. Psychological Review, 74, 392-409.

Raven, J. C. (1962). Advanced progressive matrices. Set II. London: H. K. Lewis.

Saccuzzo, D. P., & Larson, G. E. (1987). Analysis of test-retest reliability for a battery of cognitive speed tests (NPRDC TN 88-10). San Diego: Navy Personnel Research and Development Center.

U235035

SUPERINTENDENT
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5100

0142